CLAIMS:

Claims

1. An automated identification methodology for assembling document related hyperlinked pages comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page potentially part of the document; and,

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled.

2. The method of claim 1 wherein the page-level link analysis includes retrieval of referenced pages.

3. The method of claim 1 wherein the page-level link analysis includes examination of contextual clues.

4. The method of claim 3 wherein the contextual clue is a particular class of content item associated with the hyperlink.

5. The method of claim 4 wherein the class of content item is a class of text.

6. The method of claim 5 wherein the class of text is a directional word or phrase.

7. The method of claim 4 wherein the class of content item is a class of image.

8. The method of claim 7 wherein the class of image is an image containing a directional symbol.

9. The method of claim 4 wherein a textual clue is obtained for the image.

10. The method of claim 1 wherein the page-level link analysis includes the identification of progression links.

11. The method of claim 1 wherein the identification of hyperlinks includes determining the similarity of a hyperlink destination to the current page location.

12. The method of claim 1 wherein the identification of hyperlinks includes determining the similarity of hyperlink destination to that of other hyperlinks within the page.

13. The method of claim 1 wherein the page-level link analysis includes the identification of tables of contents.

14. The method of claim 13 wherein the identification is indicated by the presence of at least one other hyperlink nearby within the page description.

15. The method of claim 14 wherein the identification includes determining the similarity of the hyperlink destination to that of other hyperlinks within the page.

16. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

identifying possible progression links, and;

identifying possible table of content links;

and,

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled.

17. The method of claim 16 wherein the page-level link analysis includes examination of contextual clues.

18. The method of claim 17 wherein the contextual clue is a particular class of content item associated with the hyperlink.

19. The method of claim 18 wherein the class of content item is a class of text.

20. The method of claim 19 wherein the class of text is a directional word or phrase.

21. The method of claim 18 wherein the class of content item is a class of image.

22. The method of claim 21 wherein the class of image is an image containing a directional symbol.

23. The method of claim 18 wherein a textual clue is obtained for the image.

24. The method of claim 16 wherein the identifying of table of content links includes detecting the presence of at least one other hyperlink nearby with the page description.

25. The method of claim 16 wherein the page-level link analysis includes determining the similarity of the hyperlink destination to that of other hyperlinks with the page.

26. The method of claim 16 wherein the page-level link analysis includes determining the similarity of the hyperlink destination to the location of the current page.

27. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

identifying possible progression links;

identifying possible table of content links, and;

examining the possible progression links and the possible table of content links for common characteristics; and,

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled.

28. The method of claim 27 wherein the page-level link analysis includes examination of contextual clues.

29. The method of claim 28 wherein the contextual clue is a particular class of content item associated with the hyperlink.

30. The method of claim 29 wherein the class of content item is a class of text.

31. The method of claim 30 wherein the class of text is a directional word or phrase.

32. The method of claim 29 wherein the class of content item is a class of image.

33. The method of claim 32 wherein the class of image is an image containing a directional symbol.

34. The method of claim 29 wherein a textual clue is obtained for the image.

35. The method of claim 27 wherein the identifying of table of content links includes detecting the presence of at least one other hyperlink nearby with the page description.

36. The method of claim 27 wherein the page-level link analysis includes determining the similarity of the hyperlink destination to that of other hyperlinks within the page.

37. The method of claim 27 wherein the page-level link analysis includes determining the similarity of the hyperlink destination to the location of the current page.